SSAST: Self-Supervised Audio Spectrogram Transformer

Year: 2021 | Citations: 343 | Authors: Yuan Gong, Cheng-I Lai, Yu-An Chung, James R. Glass

---

Abstract

Recently, neural networks based purely on self-attention, such as the Vision Transformer (ViT), have been shown to outperform deep learning models constructed with convolutional neural networks (CNNs) on various vision tasks, thus extending the success of Transformers, which were originally developed for language processing, to the vision domain. A recent study showed that a similar methodology can also be applied to the audio domain. Specifically, the Audio Spectrogram Transformer (AST) achieves state-of-the-art results on various audio classification benchmarks. However, pure Transformer models tend to require more training data compared to CNNs, and the success of the AST relies on supervised pretraining that requires a large amount of labeled data and a complex training pipeline, thus limiting the practical usage of AST. This paper focuses on audio and speech classification, and aims to reduce the need for large amounts of labeled data for the AST by leveraging self-supervised learning using unlabeled data. Specifically, we propose to pretrain the AST model with joint discriminative and generative masked spectrogram patch modeling (MSPM) using unlabeled audio from AudioSet and Librispeech. We evaluate our pretrained models on both audio and speech classification tasks including audio event classification, keyword spotting, emotion recognition, and speaker identification. The proposed self-supervised framework significantly boosts AST performance on all tasks, with an average improvement of 60.9%, leading to similar or even better results than a supervised pretrained AST. To the best of our knowledge, it is the first patch-based self-supervised learning framework in the audio and speech domain, and also the first self-supervised learning framework for AST.